IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Koji Sera
September 21, 2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

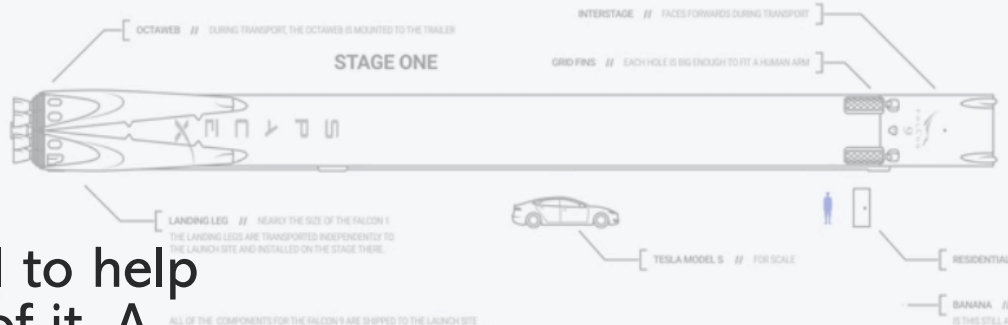
Executive Summary

- Summary of methodologies
  - Data Collection: Data of Falcon 9 launch results was collected with REST API and web scraping method
  - Data Wrangling: The obtained data included NaN. They were removed as part of data process
  - Data Analysis: Three types of data analysis was conducted: Explorative Data Analysis, Interactive Folium & Plotly Dash data visualization, and Predictive Analysis
- Summary of all results
  - Data Visualization
  - SQL
  - Machine Learning

Introduction

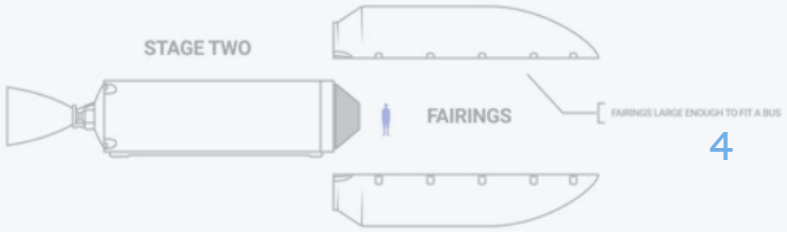
- Project background and context:

Space Y plans to launch a space rocket. As part of the plan, Space Y uses Falcon 9 launch data of Space X. The data will be cleaned and analyzed to help Space Y make a decision if they reuse the stage one or not, which is much of rocket launch cost.



- Problems to solve:

Project team plans to establish a model to help successful landing of stage one and reuse of it. A model is to be established with landing outcome as target variable and all others as independent variables.



Section 1

Methodology

Methodology

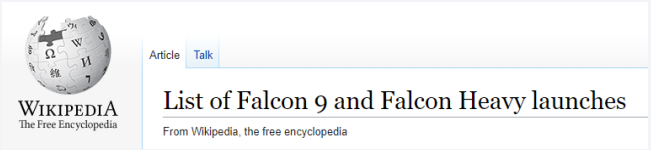
Executive Summary

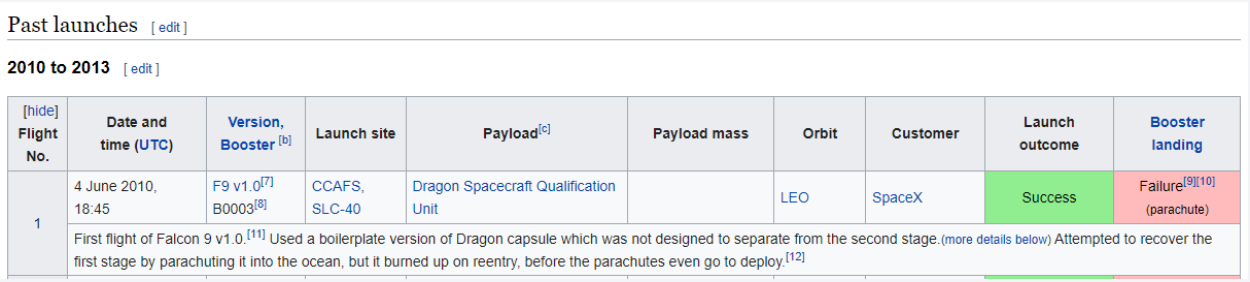
- Data collection methodology
  - Falcon 9 launch data was collected through REST API and web scraping
- Data wrangling
  - Response to one variable for landing outcome was categorized into eight (True ASDS, None None, True RTLS, False ASDS, True Ocean, False Ocean, None ASDS, False RTLS). Each response was sorted into 1 (successful landing) or 0 (unsuccessful landing).
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Four different prediction models (Logistic Regression, Support Vector Machine, Decision Tree, and K nearest neighbors) were tested for accurate outcome prediction.

Data Collection

Falcon 9 launch data was obtained through web scraping and REST API.

- Web scraping: Wikipedia shows SpaceX. The current project extracted HTML table of launch data. The table is to be converted to a Panda data frame. https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches





Past launches [edit]

2010 to 2013 [edit]

[hide] Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage. (more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even go to deploy.[12]								

- REST API: This project also obtained SpaceX launch data from the site by REST API. https://api.spacexdata.com/v4/launches/past

  The obtained JSON data was converted into a Pandas data frame.

Data Collection – SpaceX API

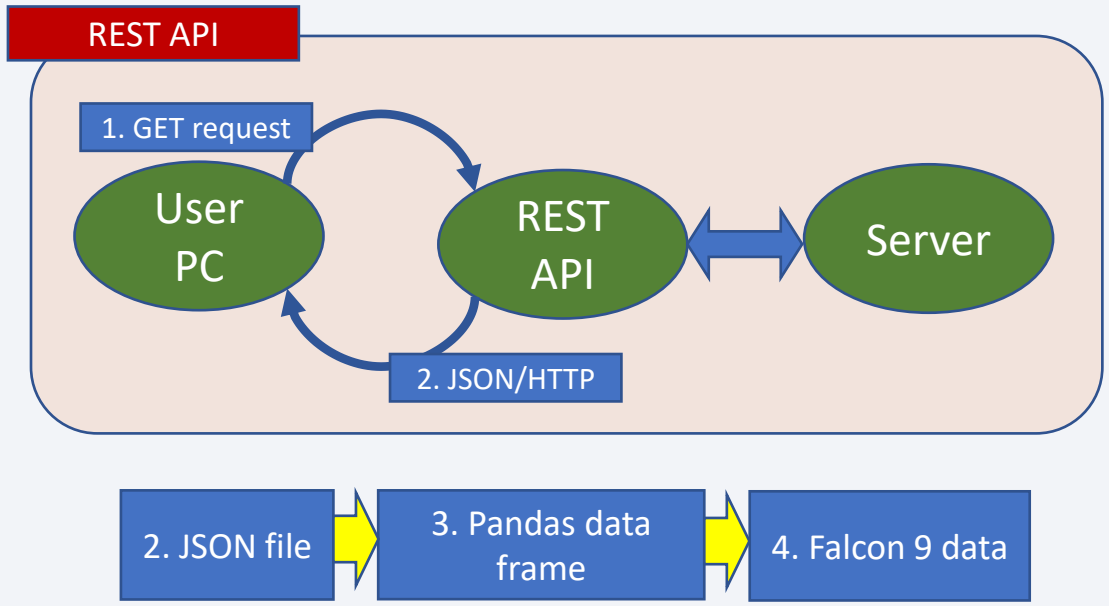
SpaceX REST API was used for Falcon 9 launch results from the following link and process:

https://api.spacexdata.com/v4/launches/past

1. Create GET request
2. Receive data in JSON format
3. Pandas data frame extraction from JSON file
4. Extraction of Falcon 9 information from the data frame.

Refer to the following link for a detailed process and the list of data frame items:

https://github.com/G-flatminor/coursera-capstone-project/blob/main/Capstone_Data%20collection%20API.ipynb



Data Collection - Scraping

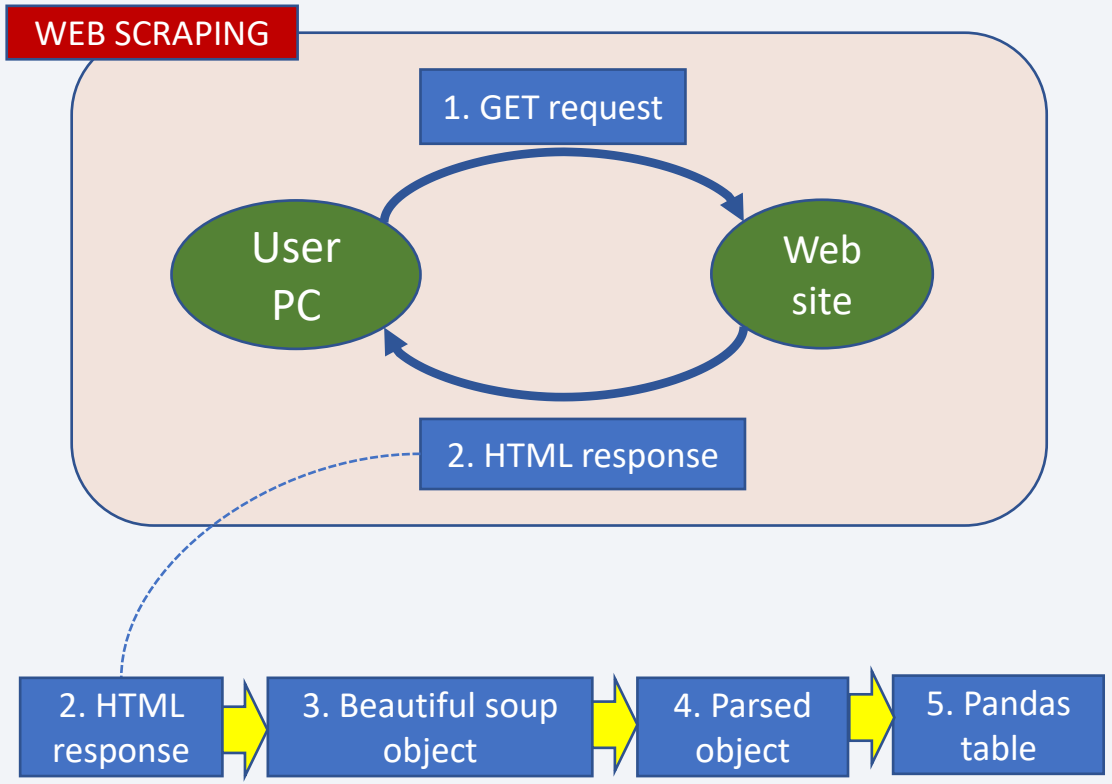
Space X data was obtained from Wikipedia by web scraping.

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

1. Create GET request
2. Scraped HTML response
3. Create Beautiful Soup object
4. Parse the object for all columns and its contents

Refer to the following link for a detailed process and the list of data frame items:

https://github.com/G-flatminor/coursera-capstone-project/blob/main/Data%20Collection%20with%20Web%20Scraping%20lab%20rev1.ipynb



Data Wrangling

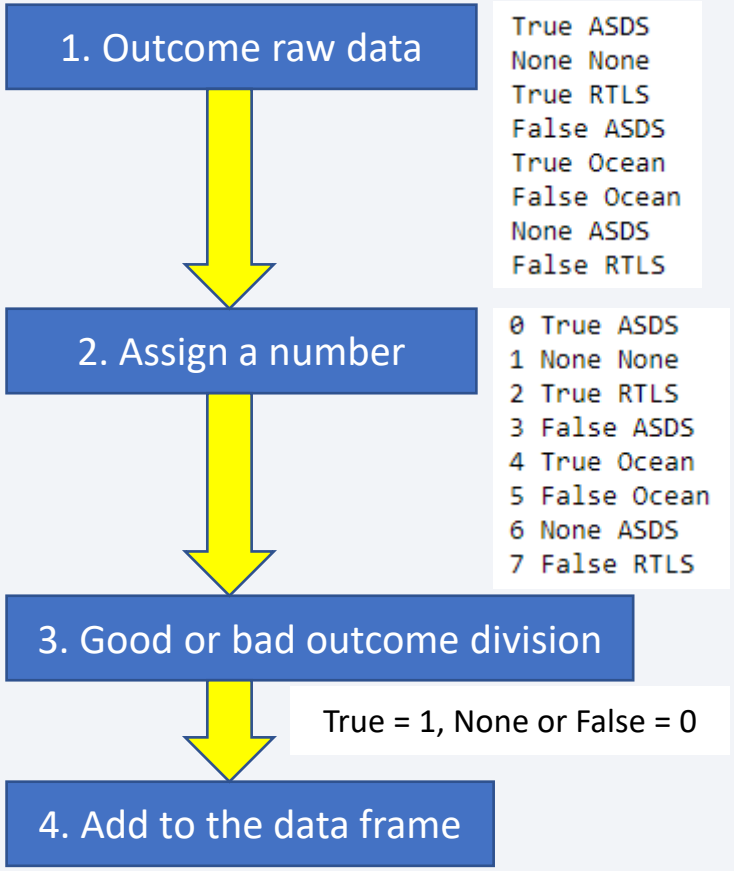
This study object is to predict the stage one landing outcome, success or failure.

1. The panda data frame shows a combination of the outcome and orbit as strings such as True ASDS, None None, True RTLS, False ASDS, etc.

2. A number was assigned to each outcome.

3. The number assigned in process 2 was labeled with either 0 (bad outcome) or 1 (good outcome).

4. Added good/bad outcome label to the original data frame.



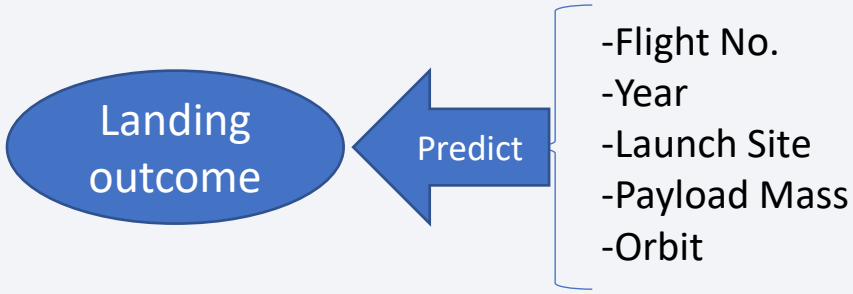
Refer to the following link for a detailed process:

https://github.com/G-flatminor/coursera-capstone-project/blob/main/Capstone_EDA%20lab%20(data%20wrangling%20with%20API%20data).ipynb

EDA with Data Visualization

- The data frame was used for data visualization.

The objective of this study is to predict success or failure of a flight based on each item (Flight No, Launch Site, Payload Mass, Orbit, and Year)



Scatter plot was applied to understand relationships of each factors against success/failure outcome.

Refer to the following link for a detailed process:

https://github.com/G-flatminor/coursera-capstone-project/blob/main/Capstone_EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

As part of EDA, SQL was used to explore:

- Launch site name
- The name of launch site which start with 'CCA'
- Total Payload Mass
- Average Payload Mass
- The date of the first flight
- Name of successful booster version
- The number of successful and failure outcome
- Name of booster version which carries the max payload
- List of failed landing outcome in drone ship in 2015
- Rank of the count of landing outcomes

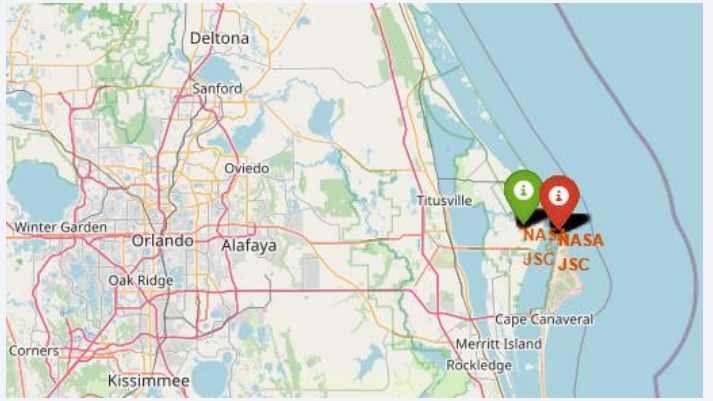
Refer to the following link for a detailed process:

https://github.com/G-flatminor/coursera-capstone-project/blob/main/Capstone_EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Falcon 9 launch sites were mapped. The mapped sites were marked with name and success/failure launch result to help understand if the geological factors influence the launch outcomes.

The mapped sites also help geological requirements for a launch site (proximity to railway and coast): the closer to railway and coast the sites are, the cheaper the cost will be.



Refer to the following link for a detailed process:

- https://github.com/G-flatminor/coursera-capstone-project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

Plotly Dash shows:

1. Ratio of usage of 4 launch sites and overall success/failure plot of each flight.
2. Success/failure ratio at each launch site and success/failure plot of each payload mass.

Plotly of success/failure ratio at each site along with payload mass shows payload mass range related to successful launch.

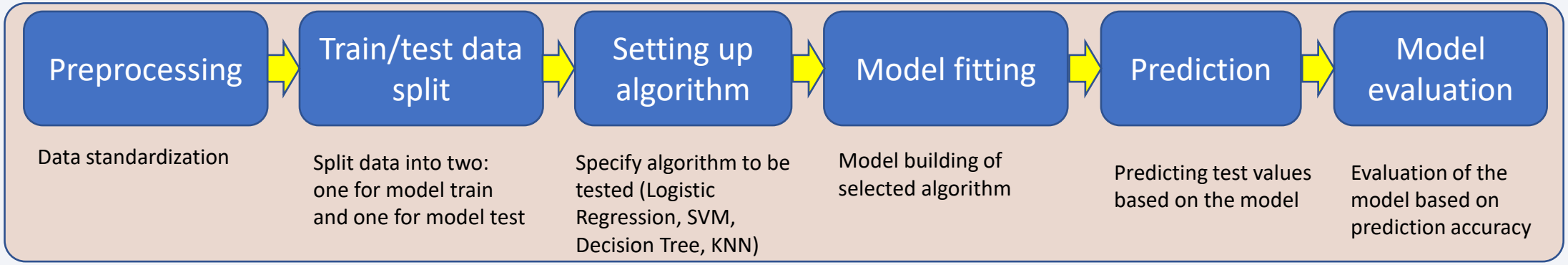
Refer to the following link for a detailed process:

- https://github.com/G-flatminor/coursera-capstone-project/blob/master/Capstone%20Plotly.ipynb

Predictive Analysis (Classification)

The study tries to conduct a predictive analysis. Four different algorithms were tested: Logistic Regression, Support Vector Machine, Decision Tree, K-nearest neighbor.

The model evaluation flow is shown below:



Refer to the following link for a detailed process:

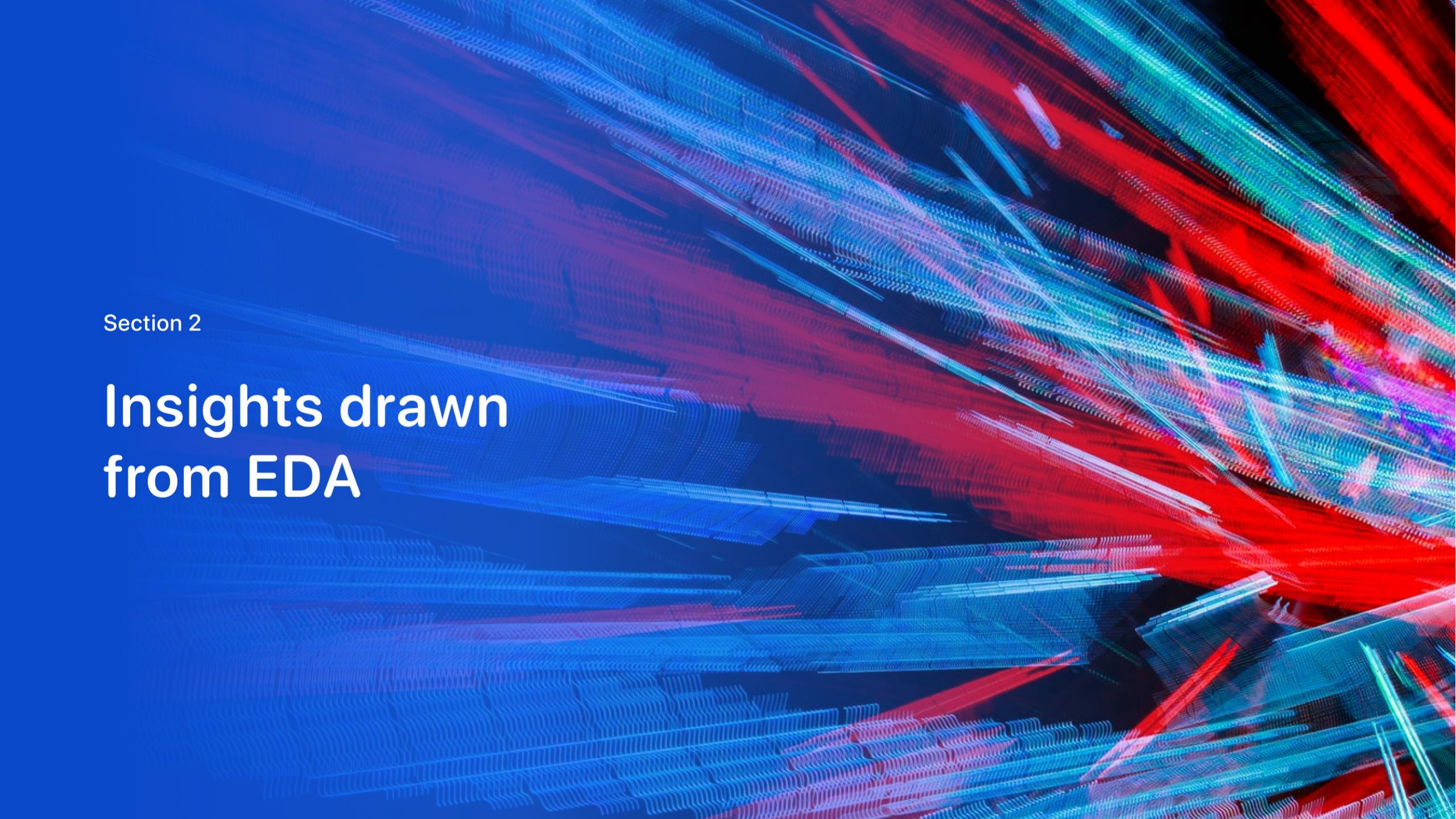
https://github.com/G-flatminor/coursera-capstone-project/blob/main/Capstone_Machine%20Learning%20Prediction%20Lab.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

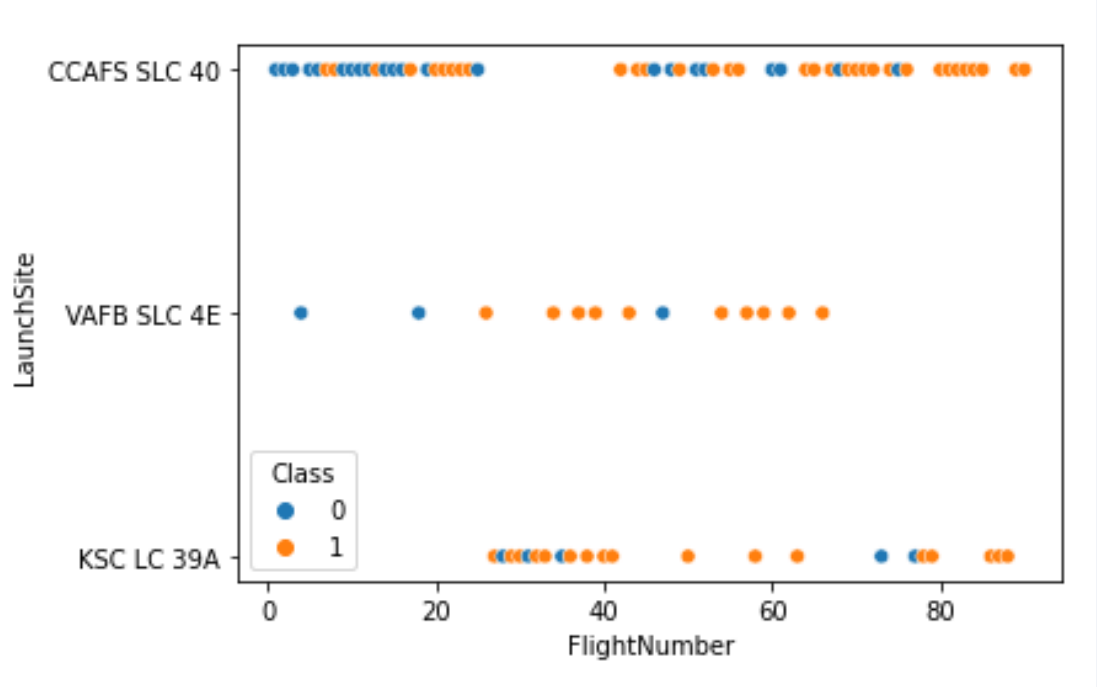
Insights drawn from EDA



Flight Number vs. Launch Site

A scatter plot is drown to show a relationship between the number of flights and launch locations along with success/failure distinction (0 = failure, 1 = success).

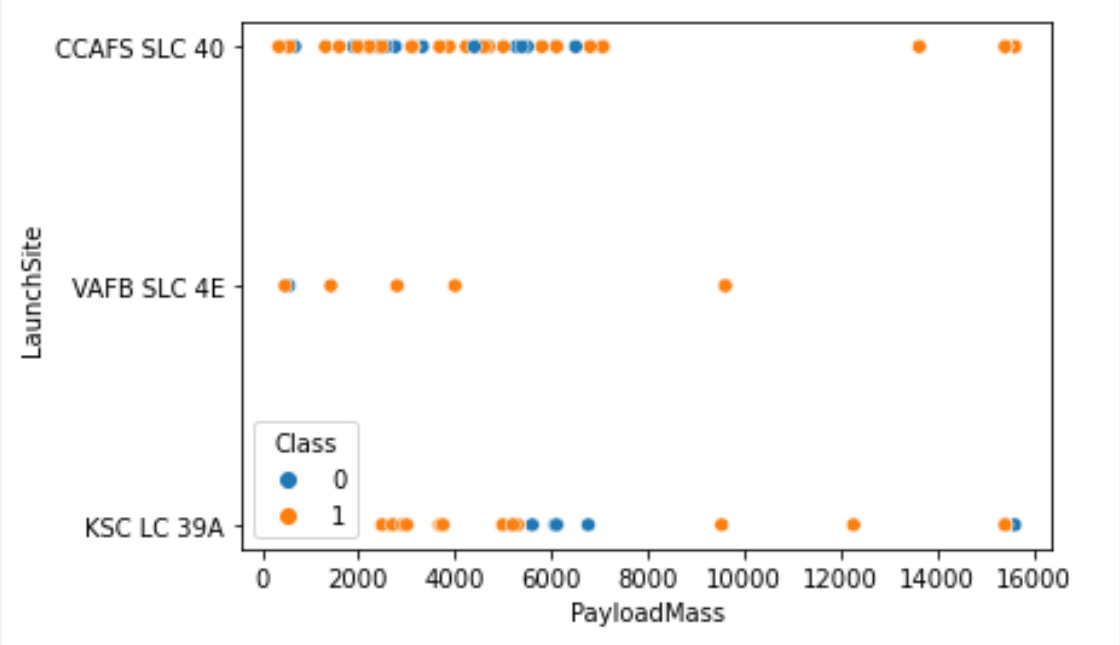
- CCAFS SLC 40 has been used the most, from the very beginning of Falcon 9 testing.
- KSC LC 39A was used the most between about flight 25 and 40 while CCAFS SLC 40 was not used at all.
- Usage of VAFB SLC 4E is rather sporadic.



Payload vs. Launch Site

Scatter plot is made to show relationship between the flight No. and Payload Mass on the right.

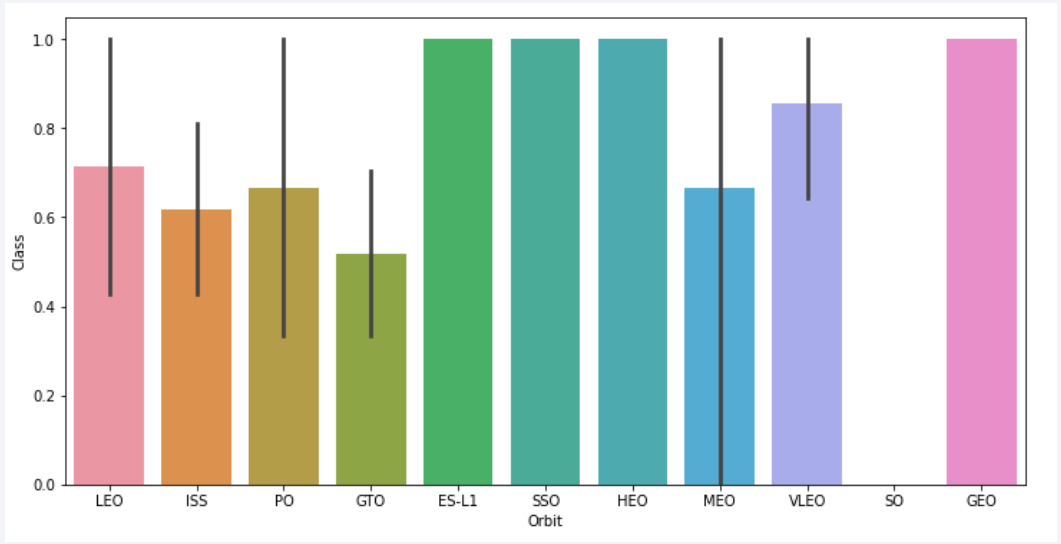
- CCAFS SLC 40 has been used the most for the range of payload mass from 0 to 8,000 kgs.
- KSC LC 39A is used the second most.
- Over 8000 kgs payload mass, those 2 locations are used almost equarly.



Success Rate vs. Orbit Type

Success rate of landing outcome by orbit types is shown on the right.

- Success rate of ES-L1, SSO, HEO, and GEO is 100%, but the number of flights are very low (1 to 5).
- The flight to SO is only one and it's a failure outcome.
- The success ratio for all other occurrences is between 50 and 90 %. VLEO shows the highest success rate (85.71%)

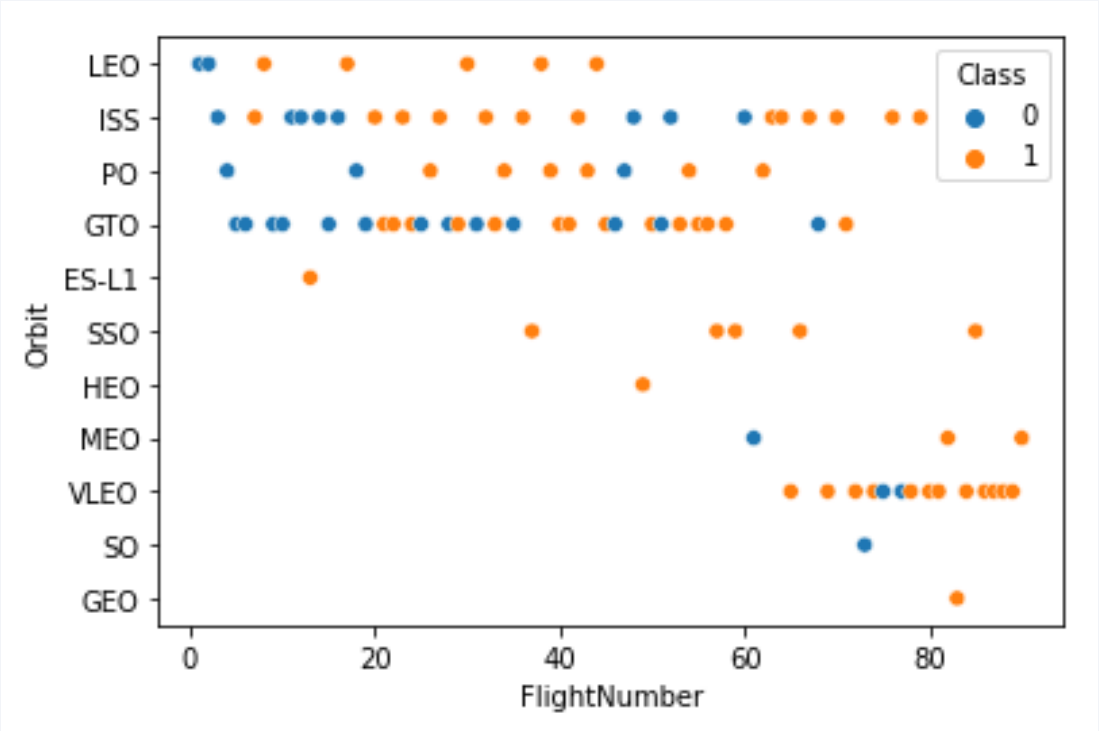


LEO: Low Earth orbit with an altitude of 500 km or less
ISS: International Space Station. multinational space station (408 km)
PO: One type of satellites with an altitude between 700 and 800 km
GTO: Geosynchronous orbit with 35786 km from Earth's equator
ES-L1: Point that cancel out the gravity between the sun and the earth
SSO: Orbit: Sun-synchronous orbit with an altitude between 600 and 900 km
HEO: Highly elliptical orbit with an altitude is between less than 1000 km and over 35756 km.
MEO: Geocentric orbits ranging from 2000 km to 35786 km
VLEO: Very low earth orbit with an altitude less than 450 km
SO: A sun-synchronous orbit with an altitude between 600 and 900 km
GEO: A circular geosynchronous orbit 35786 km above earth's equator

Flight Number vs. Orbit Type

A scatter plot on the right shows the relationship between the flight number and launch sites along with success/failure outcome.

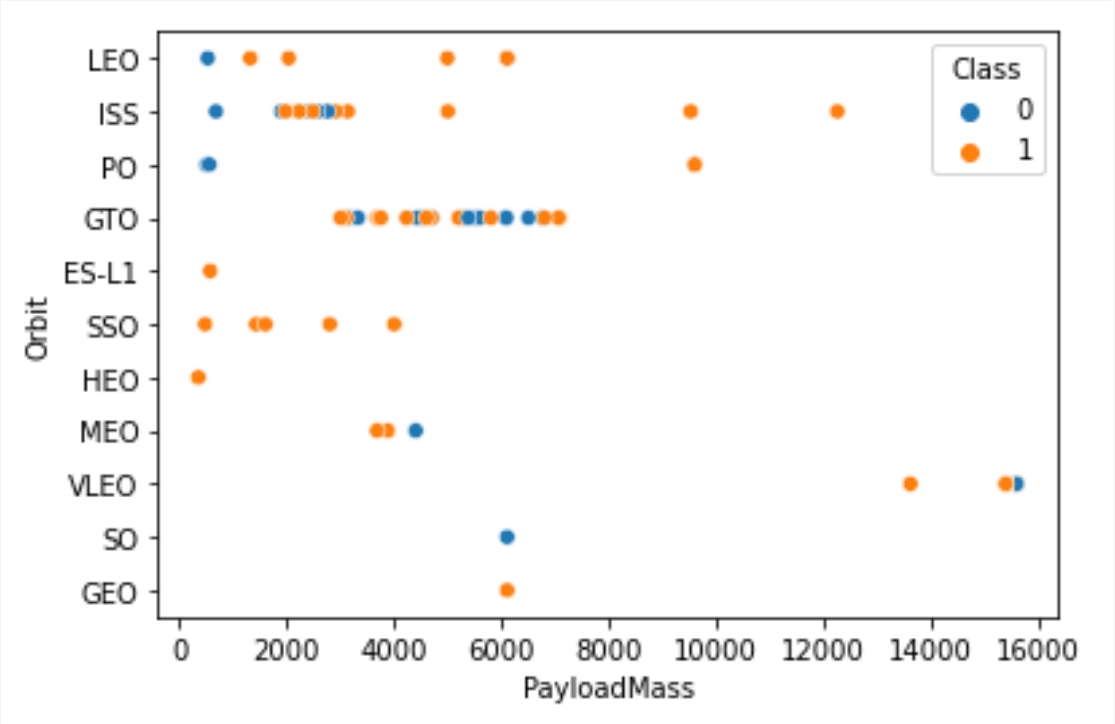
- Launch occurrences depend on the destination orbit: occurrences for ISS, GTO, and VLEO are more than all other orbits.
- Launches for ISS and GTO (and LEO) have been done since the beginning stage.
- The flights for VLEO started after the flight No. 60.



Payload vs. Orbit Type

Scatter plot on the right shows relationship between Payload Mass and Orbit type.

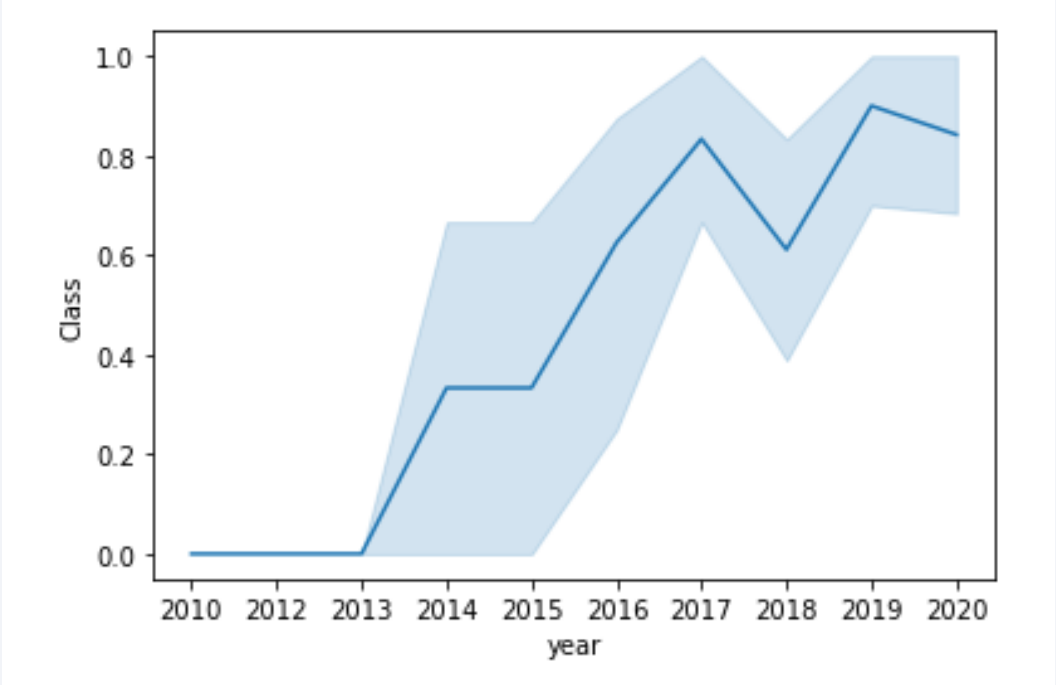
- Payload Mass range is specified based on the destination Orbit.
- Flights for ISS had mainly 2000 to 3000 kgs.
- Flights for GTO are ranged from 3000 to 7000 kgs.
- Ones for MEO are around 4000 kgs.
- Ones for VLEO are from 13500 to 16000 kgs.



Launch Success Yearly Trend

The annual launch success rate is plotted on the right.

- The early stage of flights success rate is low, up to 35% (from 2010 to 2015).
- From 2015 to 2017, the flight success rate is greatly improved from around 35% to over 80%.
- In 2018, the success rate dropped around 60%.
- 2019 and 2020, the success rate was improved over 80% again.



All Launch Site Names

To find unique launch sites, the following query was run;

```
%sql select distinct launch_site from spacextbl
```

And the following is the query result;

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

CCAFS LC-40, CCAFS SLC-40, and KSC LC-39A are located in Florida, while VAFB SLC-4E is in California.

Launch Site Names Begin with 'CCA'

Query to find 5 launch site records starting with ‘CCA’ is as follows;

```
%sql select launch_site from spacextbl where launch_site like 'CCA%' limit 5;
```

The result of the query is shown below.

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

The result shows only “CCAFS LC-40”, but “CCAFS SLC-40” also meets a criterion “where launch_site like ‘CCA%’.

Total Payload Mass

Total payload mass carried by boosters from NASA is calculated with the following query:

```
%sql select sum(payload_mass__kg_) from spacextbl where customer like 'NASA%';
```

The result is as follows:

1
99980

Total of 99980 kgs of payload mass were carried by boosters from NASA.

Average Payload Mass by F9 v1.1

Total payload mass carried by booster version F9 v1.1 is calculated with the following query:

```
%sql select avg(payload_mass__kg_) from spacextbl where booster_version = 'F9 v1.1';
```

The result is as follows:

1
2928

2928 kgs of payload mass were carried by F9 v1.1 booster.

First Successful Ground Landing Date

The first successful landing outcome date in ground pad is found out with the following query:

```
%sql select min(date) from spacextbl where landing__outcome like 'Success%';
```

The query result is below.

1
2015-12-22

The date of the first successful landing outcome in ground is Dec 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

The following query was run to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 kgs.

```
%sql select booster_version from spacextbl where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

The query result is here:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

These 4 boosters have the result of successful landing on drone ship with payload between 4000 and 6000 kgs.

Total Number of Successful and Failure Mission Outcomes

The following query was run to count the total number of successful and failure mission outcomes:

```
%sql select mission_outcome, count(*) from spacextbl group by mission_outcome;
```

The query result is here:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Failure missions were counted as 1, and the bottom item is counted as success. Final result is 100 success vs 1 failure.

Boosters Carried Maximum Payload

The following query was run to find out the names of the booster which have carried the maximum payload mass:

```
%sql select booster_version from spacextbl where payload_mass__kg_ =(select max(payload_mass__kg_) from spacextbl);
```

The query result is here (total 12 boosters carried the maximum payload):

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

The following query was run to find out the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;

```
%sql select landing__outcome, booster_version, launch_site from spacextbl where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

The result is here:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

These two occasions are the failure results which meet the conditions above.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

The following query was run to list the counts of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql select landing__outcome, count(landing__outcome) from spacextbl where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by count(landing__outcome)
```

The query result is on the right.

There are 10 occasions counted as “No attempt”. Besides those cases, success rate is over 50 %.

landing__outcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

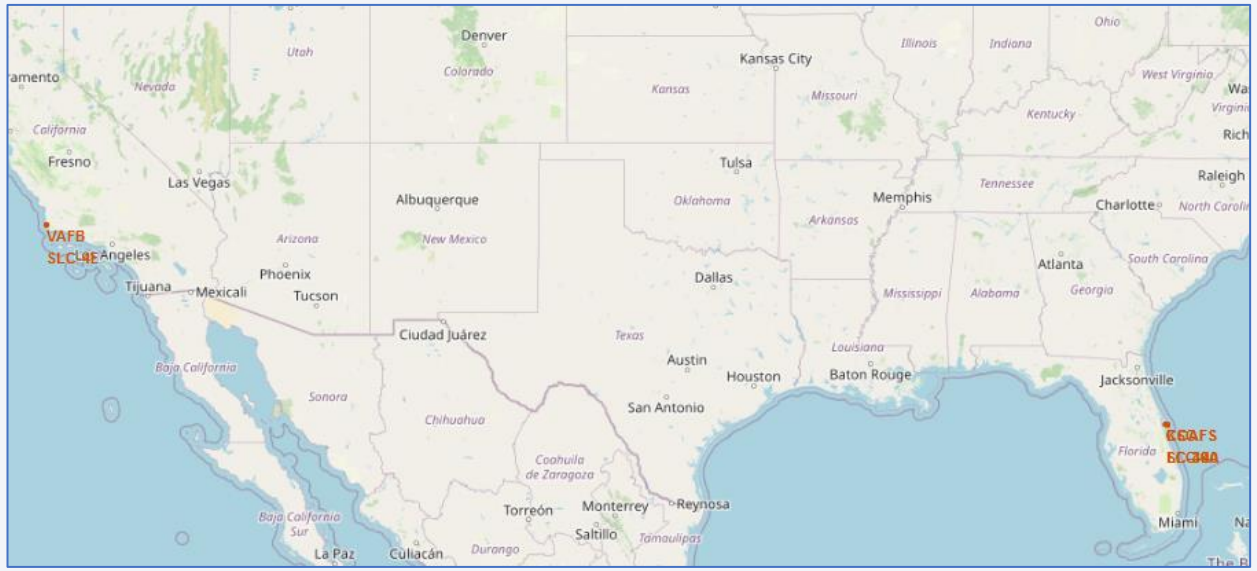
Launch Sites Proximities Analysis

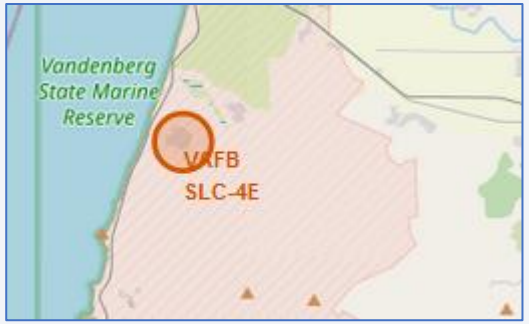
Falcon 9 launch location map

All launch sites were mapped with mark.

The top one shows in a big picture.

The bottom ones show in a closer map, and this shows all sites are close to railway and coast.





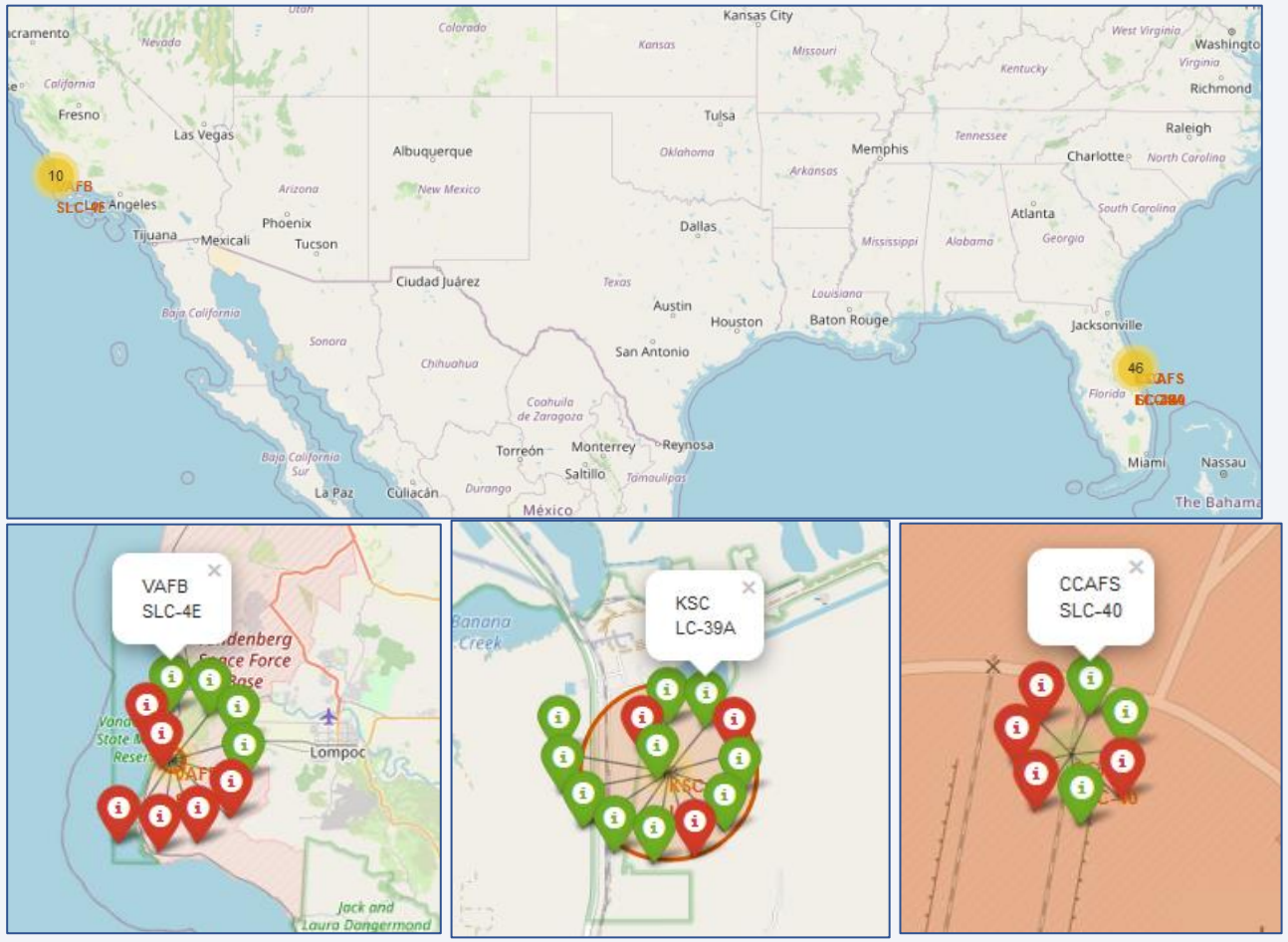
Falcon 9 launch location map with launch outcome

All launch sites were mapped with mark for success/failure launches.

The top one shows in a big picture.

The bottom ones show in a closer map, and this shows green or red marking.

The red marking means a failure outcome while the green one indicates a successful outcome.



Proximities to important sites (1/3)

Next three slides show distance between two important sites, railway and coast.

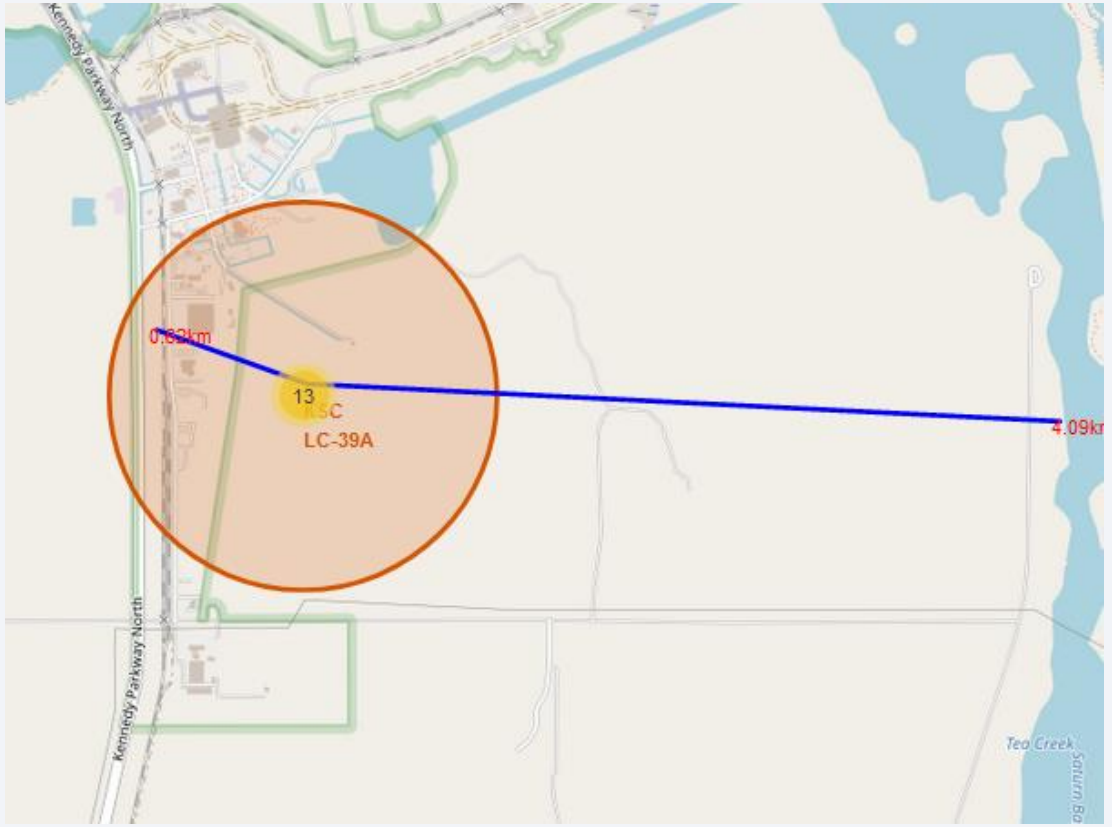
Two connecting lines show proximity between the site and railway and coast.

The image on the right:

KSC LC-39A to railway: 0.62 km

KSC LC-39A to coast: 4.09 km

While the distance to railway is very close, the distance to the coast is rather far.



Proximities to important sites (2/3)

Distance from CCAFS SLC-40/LC-40 to railway and coast is shown on the right.

To railway: 1.23 km

To coast: 0.89 km

The distance to the coast is closer than the one to railway.

Proximities to important sites

Distance from VAFB SLC-4E to railway and coast is shown on the right.

To railway: 1.25 km

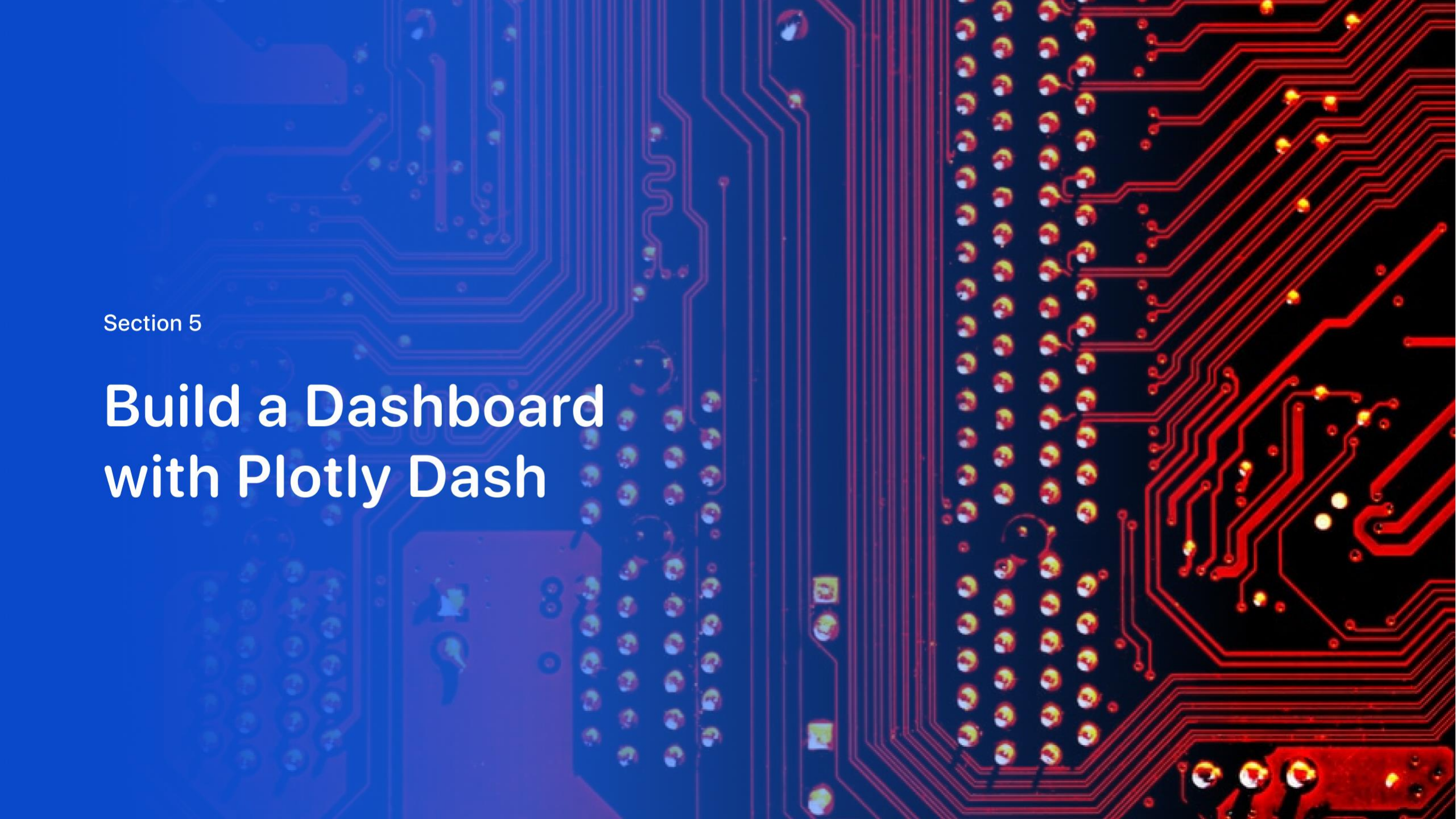
To coast: 1.36 km

The distance to the coast is similar to the one to railway.

Out of these 3, CCAFS SLC-40/LC-40 seems most cost efficient and safe

Section 5

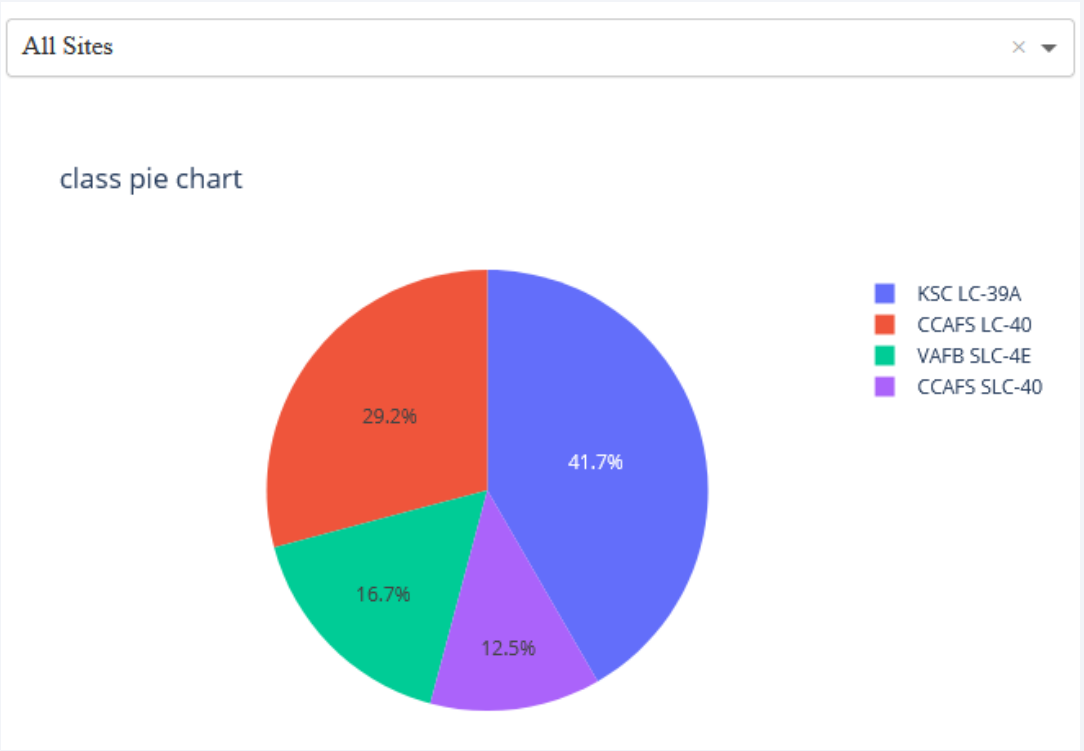
Build a Dashboard with Plotly Dash



Successful launch by launch site

The pie chart on the right shows which launch site has the most successful outcomes ratio.

Launch Site	Success Rate	No of successful outcomes
KSC LC-39A	41.7 %	10
CCAFS LC-40	29.2 %	7
VAFB SLC-4E	16.7 %	4
CCAFS SLC-40	12.5 %	3

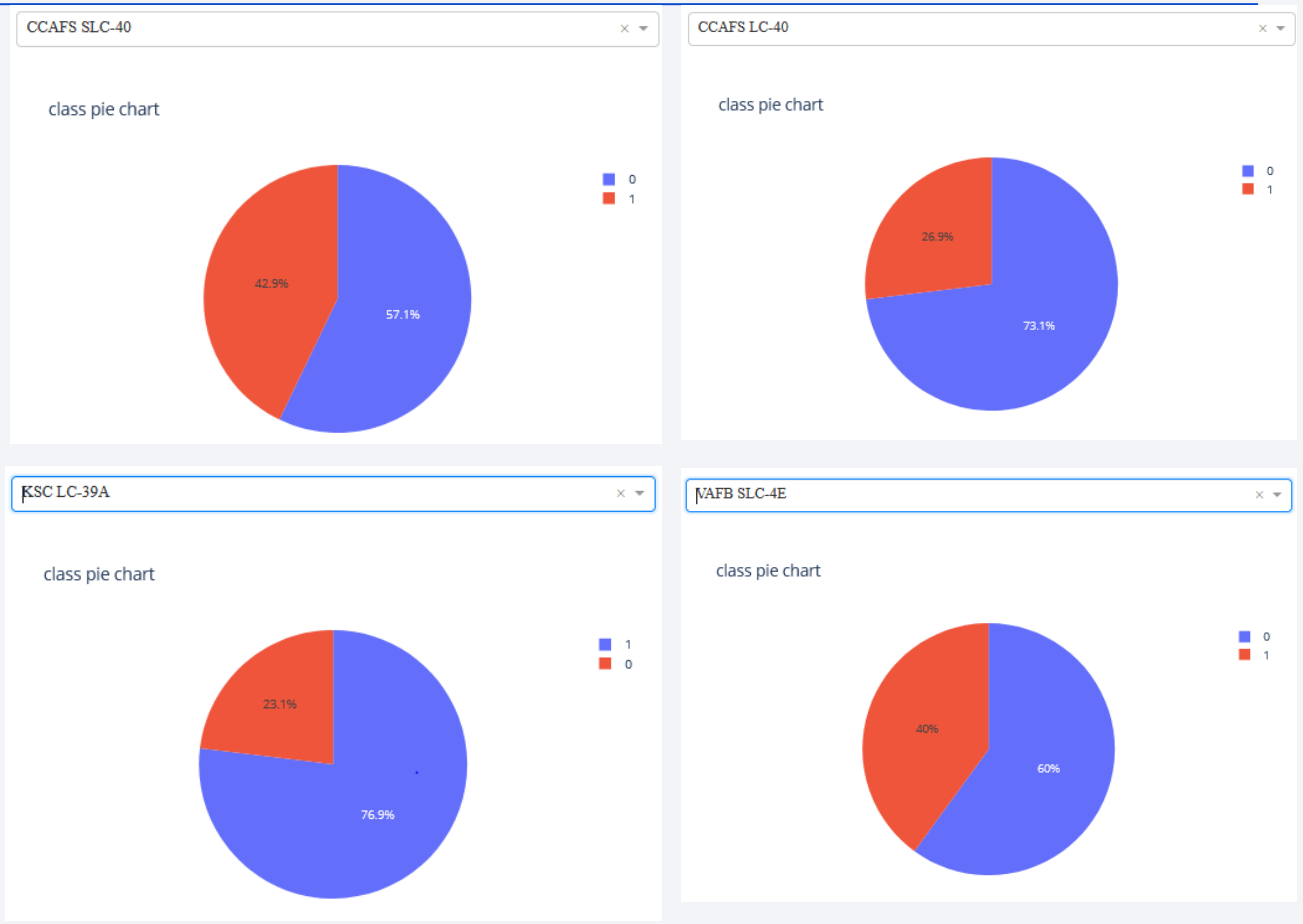


- The chart shows KSC LC-39A has experienced the most successful outcome ratio.
- KSC L-39As' success is also experienced at the highest success ratio.

Successful outcome ratio by launch site

The pie charts on the right show the successful outcome ratio by launch site.

Launch Site	Success Rate	Successful outcomes
KSC LC-39A	76.9 %	10
CCAFS LC-40	26.9 %	7
VAFB SLC-4E	40.0 %	4
CCAFS SLC-40	42.9 %	3



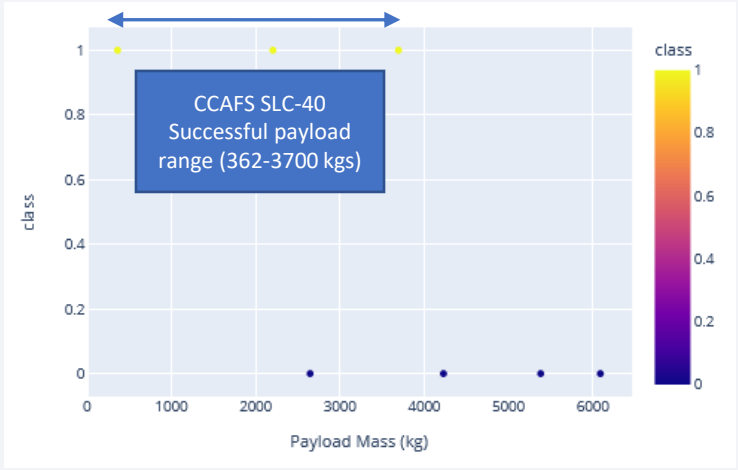
KSC LC-39A has both the highest success rate and the most successful outcomes. CCAFS LC-40 has the second successful outcomes, but its success rate is the lowest out of the four sites.

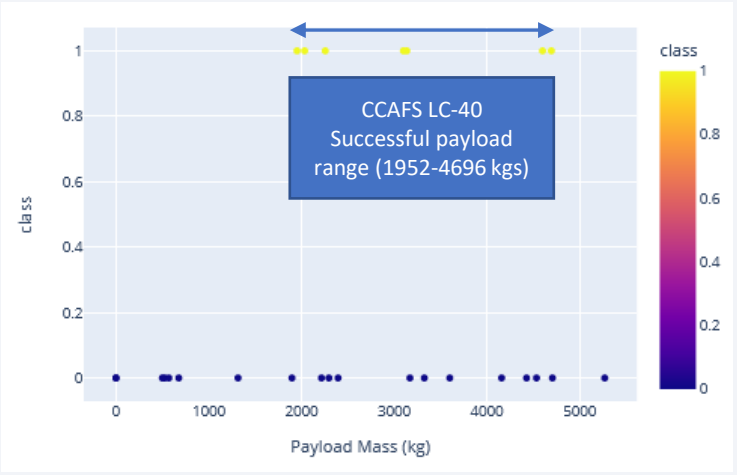
Scatter plot of Payload vs Launch Outcome by Site

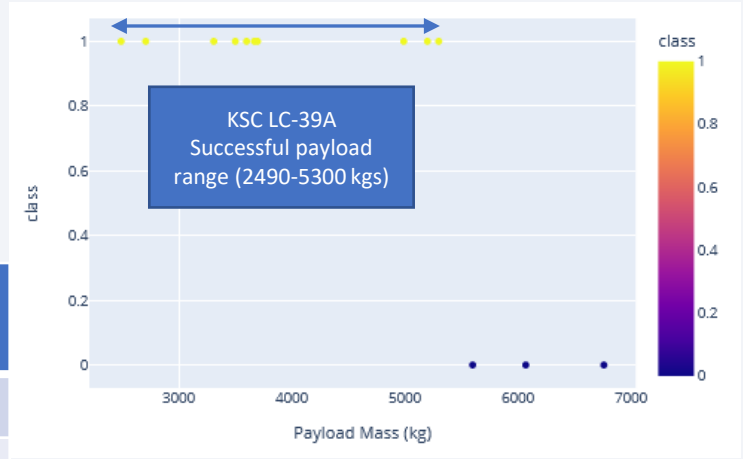
Scatter plots on the right show outcome of each launch with its payload mass at each site.

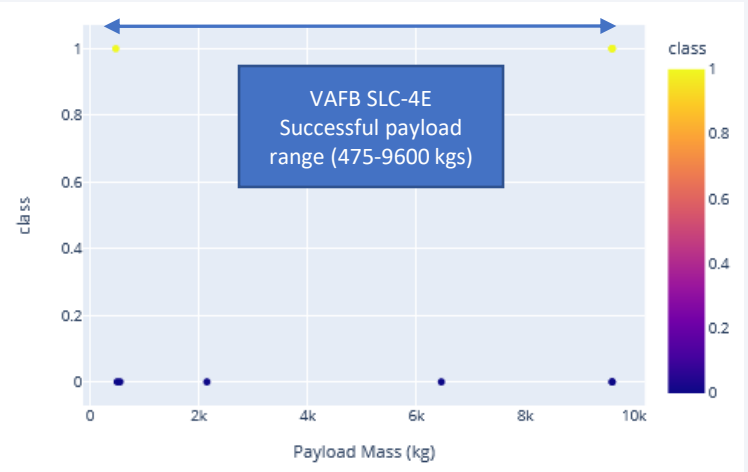
- Successful outcomes are ranged from 362 kgs to 9600 kgs.
- Unsuccessful outcomes are scattered in all ranges.
- It seems no direct connection between launch outcome and payload mass.

Launch Site	Successful outcomes	Payload Range of successful launch
KSC LC-39A	10	2490-5300 kgs
CCAFS LC-40	7	1952-4696 kgs
VAFB SLC-4E	4	475-9600 kgs
CCAFS SLC-40	3	362-3700 kgs









Section 6

Predictive Analysis (Classification)

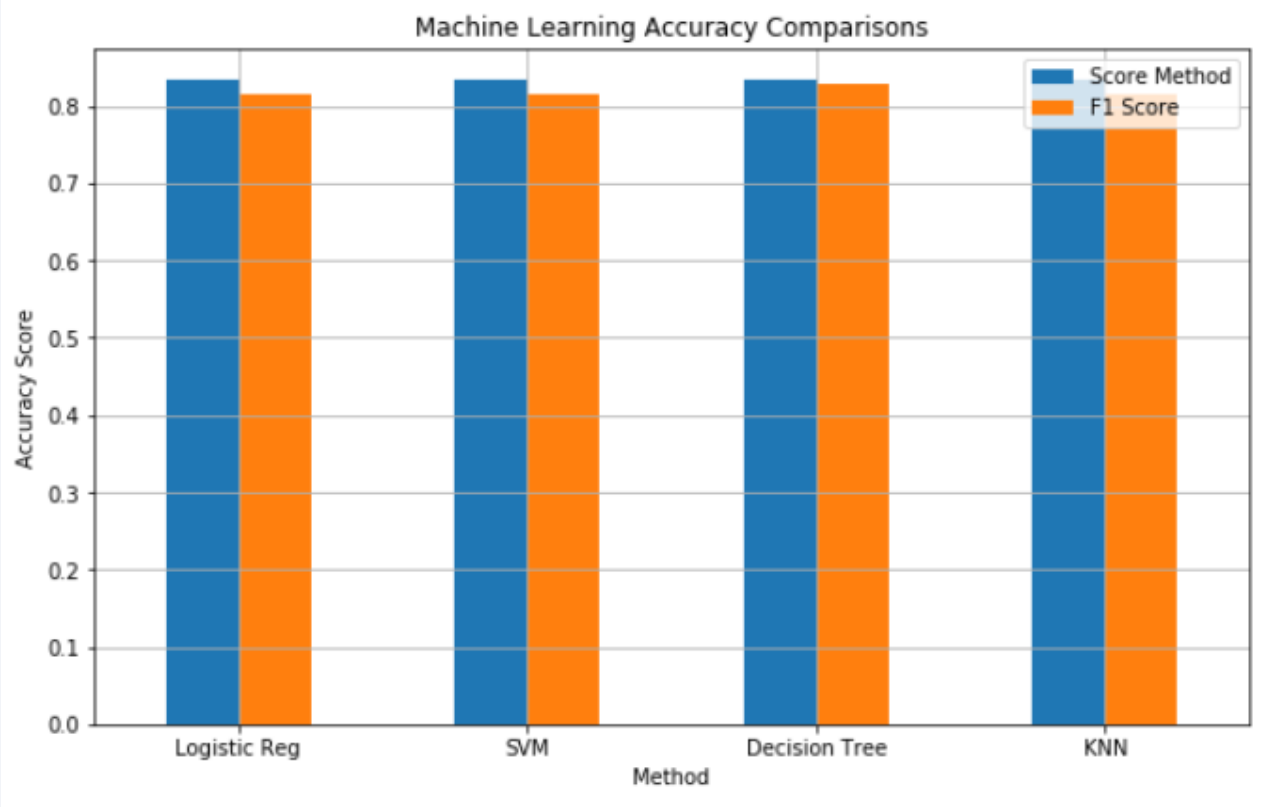
Classification Accuracy

This study try to establish a model to predict Falcon 9 launch outcome (Logistic Regression, Support Vector Machine, Decision Tree, and K-nearest neighbors)

For all models, Score Method and F1 score are calculated.

The figure on the right shows the two types of accuracy score for all models.

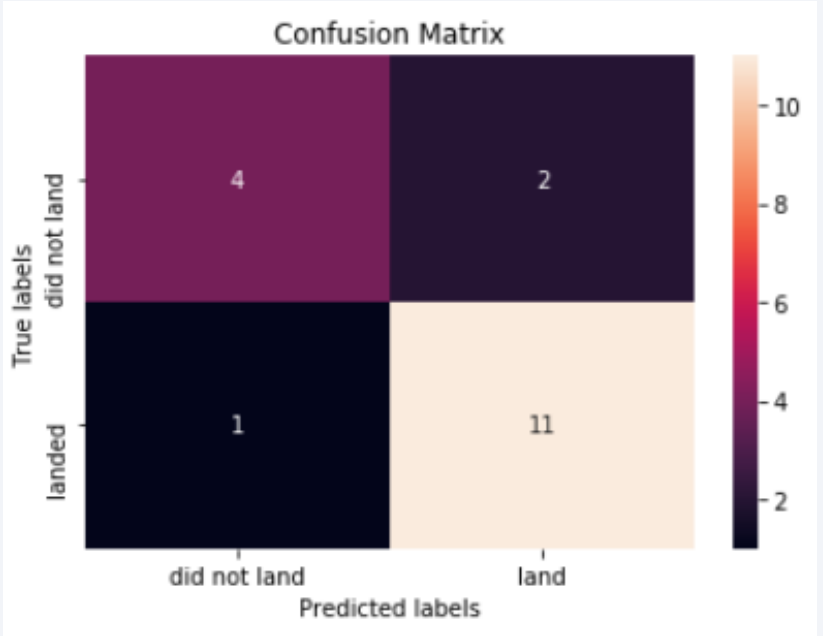
- F1 score of Decision Tree is higher than all other models.
- All models show the same level of Score Method, 0.8333.

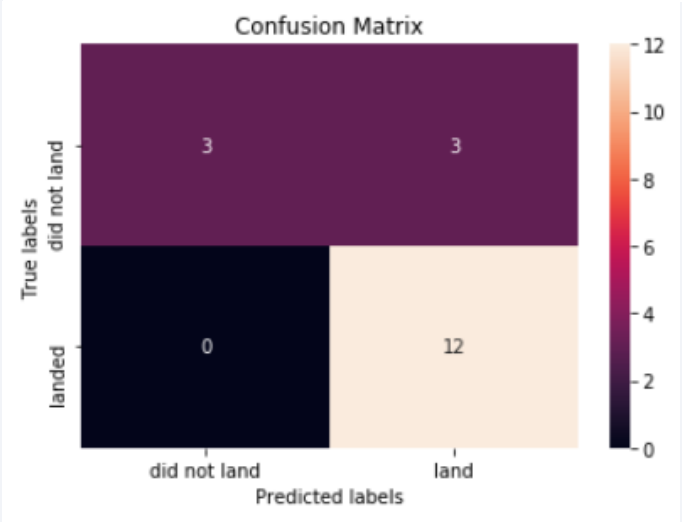


Confusion Matrix

Confusion matrix of Decision Tree method is shown on the top right.

- The model incorrectly predicted 3 cases.
  - It predicted 2 cases of the stage one landed, but they didn't land in reality.
  - The model predicted one case of the stage one didn't land, but it landed successfully.
  - Other models also predicted 3 cases incorrectly, as shown on the bottom right.





Conclusions

- This study shows the current model can predict success/failure outcome of each flight at accuracy of 82-83%.
- All prediction models show the same level of prediction accuracy.
- This study used 90 flights. 72 cases were used for model training, and 18 cases were for model testing. This size of cases may have impacted the result of machine learning modelling. In other words, 4 machine learning models didn't differ in accuracy level because of small size of cases.
- For successful outcome cases, total cost can be added for further analysis.
- Refer to the following link for a detailed process:

https://github.com/G-flatminor/coursera-capstone-project/blob/main/Capstone_Machine%20Learning%20Prediction%20Lab.ipynb

Thank you!